DOCUMENT RESUME

ED 449 231 TM 032 376

AUTHOR Smith, Tracy D.; McMillan, Bradley F.
TITLE A Primer of Model Fit Indices in Structural Equation Modeling.
PUB DATE 2001-02-00
NOTE 15p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Chi Square; *Goodness of Fit; *Structural Equation Models

ABSTRACT

This paper reviews the theoretical background, optimal levels, strengths, weaknesses, and additional considerations of the most frequently used structural equation modeling (SEM) fit statistics in an effort to enable researchers to make better, more informative judgments regarding their models. Fit indices evaluate model fit for the data being examined. Models demonstrate overall fit and the local fit of individual parameters. Some of the commonly used fit indices discussed are: (1) chi squared; (2) goodness of fit and adjusted goodness of fit indices; (3) normed fit and nonnormed fit; (4) comparative fit index; and (5) root mean square error of approximation. Researchers using SEM must determine whether they are interested in testing the null hypothesis, absolute fit, or incremental fit, and they should be aware of the shortcomings of different fit statistics and how the model may lessen the applicability of specific fit statistics. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running head: FIT STATISTICS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

T. D. Smith

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

ED 449 231

A Primer of Model Fit Indices in Structural Equation Modeling

Tracy D. Smith, M.S.

University of Southern Mississippi

&

Bradley F. McMillan, Ph.D.

Pine Grove Recovery Center

TM032376

Paper presented at the annual meeting of the Southwest Educational Research Association, February 1-3, 2001, New Orleans.

BEST COPY AVAILABLE

A Primer to Model Fit in Structural Equation Modeling

Researchers without extensive statistical training may find statistical methods such as structural equation modeling daunting, to say the least. Despite the availability of numerous methods of assessing model fit (or perhaps because of the proliferation of so many alternate methods), researchers may have difficulty assessing their model's fit to their data due to the conflicting nature of many assessment strategies. This paper is an attempt to highlight and explain various methods to assess fit in structural equation models.

The growth of computers and computer technology helped make SEM possible, while the theory behind SEM dates from before the 1960s, it didn't become a unified statistical procedure until the introduction of LISREL created by Jöreskog and Sörbom (1989). The primary purpose of SEM is to test so-called causal theories by allowing factor analysis and multiple regression type procedures to be performed simultaneously. In addition, constructs can be more readily identified because of SEM's ability to partial out measurement as part of the model. The potential value of SEM, however, still relies on the following: the status of the hypothesis relative to the theory, how well the constructs have been operationalized, and the "match" between the hypothesis and the statistical procedure used to test it.

While statistical textbooks provide appropriate numerical criteria in order to evaluate model fit they often fail to elaborate on the theory behind their creation, thus leaving researchers to believe these cut-offs are to be strictly observed without understanding their importance. This primer will address the theoretical background, optimal levels, strengths, weaknesses, and additional considerations of the most frequently used SEM fit statistics in an effort to better enable researchers to make informative judgements regarding their models.

Overview of Fit Indices

Fit indices evaluate model fit for the data being examined. They help the researcher determine which proposed model(s) best fit the data by showing how well the parameter estimates account for the observed covariances. Models demonstrate two main types of fit: overall fit and the local fit of individual parameters. Overall fit is evaluated by how well the model explained all of the data in the entire analysis. Local fit is determined by examining how specific parameters that are free for estimation may have achieved statistical significance within the model. The chi-squared statistic (χ^2) is the generally recognized fit index for assessing overall model fit. It tests the null hypothesis of no difference between the proposed model and the data structure, and good-fitting models should retain the null hypothesis (i.e., the chi-squared statistic should not be significant). However, several researchers over the years have criticized the use of χ^2 because of its shortcomings, primarily that it is a statistical significance test and – as such – is heavily impacted by sample size, making retention of the null for large sample almost impossible. As a result of these criticisms, a number of additional fit indices have been created. Chi-squared and these additional fit indices are described in more detail in the rest of this paper.

Overall Fit

Overall model fit requires consideration of several indicators. Bollen and Long (1993) argue that the first guide to measuring the adequacy of a particular model is strong substantive theory. Empirically, however, model fit is evaluated with several indices that provide different information. The three most widely-cited empirical criteria are: tests of the null hypothesis, tests of absolute fit, and tests of incremental fit. While these three areas are the most common empirical examinations of model fit, others have proposed other criteria (Tanaka, 1993). Given

all the different types of model evaluations, it is important to evaluate model fit with multiple fit indices to avoid making inaccurate assumptions of a model's fit to the data being examined.

Tests of the null hypothesis are assessed using a chi-squared statistic (Hu & Bentler, 1995). Because there are many chi-squared tests, however, it is important to know which statistic is best to use. The maximum likelihood (ML) and generalized least squares (GLS) are the two estimation theories most frequently used when determining the chi-squared statistics in most SEM analyses. Both of these tests assume multivariate normality of the data (Henson, 1999; Hu & Bentler, 1995).

Because the overall model test that is represented by the chi-squared statistic has a number of difficulties associated with it (such as the aforementioned sample size problem), researchers began to look at other means of assessing model fit (La Du & Tanaka, 1995). The earliest explorations led to the development of what are currently termed "absolute fit indices." Absolute fit indices employ as part of their computation the sample covariance matrix and the estimated population matrix as derived from the model being tested. Often other elements, such as degrees of freedom for the model tested, sample size, and/or the number of measured variables in the model are also included in computing these indices. The two best known absolute fit indices are the goodness of fit index (GFI) and adjusted goodness of fit index (AGFI). While the chi-squared statistic can also be lumped into the absolute fit index category, GFI and AGFI are more commonly discussed when talking about absolute fit.

Subsequent to the development of the absolute fit indices, other researchers developed what are currently termed "incremental or relative fit indices." Incremental fit indices require not only the two matrices used in absolute fit indices, but a third matrix as well. This third matrix is used as an aid in assessing model fit. Several of these indices have been developed, but the most

widely-used are: comparative fit index (CFI), normed fit index (NFI), and the relatively new root mean square error of approximation (RMSEA).

Besides the fit indices discussed in this paper, other indices have also been created. While some of these have dropped out of fashion or become subsumed in other fit indices (for example, TLI, IFI, and BFI), others are still being used and have their adherents (such as RMR and RNI). Given the ever-evolving nature of SEM analyses, it is likely that new methods of evaluating model fit will continue to be explored and developed. In the meantime, the following indices are likely to give beginning researchers a good place to start in evaluating their work.

Chi-Squared Test

Chi-squared is the conventional overall test of fit in structural equation modeling. The chi-squared test enjoyed substantial popularity when it was first proposed as a means of testing models (Joreskog, 1969), because it made confirmatory factor analysis free of the subjective decisions that were being employed up until the use of the chi-squared statistic.

The chi-squared test assesses the magnitude of discrepancy between the sample and the fitted covariance matrices. The chi-squared analysis of covariance matrices is at the heart of SEM and explains why SEM has frequently been called "covariance structure analysis" or "analysis of covariance matrix structures." Parameters in SEM are estimated so that the discrepancy between the sample covariance matrix and the implied covariance matrix is minimal. A statistically nonsignificant chi-squared result is considered optimal, which indicates no statistical difference between the sample and model covariance matrices. The chi-squared analysis is conducted using a t statistic. A large t statistic relative to the degrees of freedom associated with the model indicates that the model may not be a good fit to the data. While there are several different t statistics, the maximum likelihood (ML) and generalized least squares

(GLS) estimation methods are the most widely employed summary statistics for assessing the adequacy of a structural equation model. These methods assume normality of the data being examined, however, and lose their robustness when this assumption is violated (Fan & Wang, 1998).

Historically, one of the key strengths of the chi-squared analysis in SEM is that it finally gave researchers an objective way to assess model fit. Up until the proposal of the chi-squared statistic, researchers evaluated exploratory models based upon factor analytic rotation methods and subjective factor analysis techniques. Thus, chi-squared provided, for the first time, a means of evaluating models with more objective criteria. Problems associated with the chi-squared tests were recognized quite early, however. One of the main concerns of researchers is the chi-squared test's susceptibility to sample size. With small sample sizes, the chi-squared test may lack power and not be able to discriminate poor models from adequate ones. And, with larger sample sizes, trivial differences between the implied and tested models may lead to a rejection of an adequate model. Thus, the standard chi-squared may not be a good enough guide to model adequacy, because a statistically significant chi-squared value may be the result of model misspecification, the power of the test, or a violation of some technical assumptions underlying the estimation method.

GFI & AGFI

The goodness of fit (GFI) and adjusted goodness of fit (AGFI) indices were developed by Jöreskog and Sörbom (1984) as alternatives to the chi-squared statistic and its limitations. GFI was the original fit index created for LISREL, the SEM computer program developed by Jöreskog and Sörbom. GFI and AGFI essentially compare the ability of a model to reproduce the variance/covariance matrix to the ability of no model at all to do so. The AGFI adjusts the GFI

for the number of degrees of freedom expended in estimating the model parameters. Indices less than zero are treated as zero (indicating no model fit) and range up to one (indicating perfect model fit). Many researchers have established .9 as an appropriate cutoff for determining adequate model fit. More recently, however, researchers have specified the .92 or .95 levels as more adequate cutoffs for GFI & AGFI (Bollen & Long, 1993).

GFI and AGFI have the benefit of being more specific indices of fit than the chi-squared statistic and they take degrees of freedom into account and eliminate some of the problems inherent in the chi-squared statistic alone. Some research indicates that for GFI-based power analyses, however, holding null and alternative values of GFI fixed leads to decreased power as degrees of freedom increases. This can lead to more difficulty in detecting false null hypotheses. This is not a problem for AGFI analyses, as power increases as degrees of freedom increase. But for both indices, it is shown that it is problematic to establish consistently appropriate values for null and alternative hypotheses about model fit (MacCullum & Hong, 1997). Thus, sample size does impact detrimentally upon these indices in some instances.

NFI & NNFI

The normed fit index (NFI) was one of the earliest fit indices. Developed by Bentler and Bonnett in 1980, NFI is an incremental fit index. This index assesses fit by comparing the tested model with a more restricted null model in which all observed variables are assumed to be uncorrelated. Given the shortcomings that they found in NFI, however, Bentler and Bonnet (1980) proposed NNFI (the nonnormed fit index) as an alternative. NNFI involves the chi-squared/degrees of freedom ratio rather than just the simple chi-squared value found in other indices. NNFI is based on the Tucker-Lewis index (TLI, 1973), which was developed through factor analysis. Unlike NFI, NNFI can exceed the 0 to 1 range. However, NFI and NNFI are both

used in an absolute sense, where 1 equals a perfect model fit and 0 equals a complete lack of fit. An index value of .9 or above has been conventionally regarded as indicating good to excellent fit for both fit indices (Bentler & Bonnet, 1980).

NFI and NNFI are both valuable, because they are less affected by problems inherent in the use of the absolute fit indices and chi-squared analyses. In subsequent research, however, Bentler found that NFI did have some inherent weaknesses, most notably when sample sizes were small (Bentler, 1993).

CFI

The comparative fit index (CFI) was created by Bentler as another alternative to NFI (Bentler, 1993). Given that NFI has been shown to be an underestimate when small samples are used, CFI was developed. CFI is another incremental fit index, and it is based on an earlier Bentler statistic (BFI, the Bentler Fit Index) and McDonald and Marsh's (1990) relative noncentrality index (RNI). CFI is often heralded as a better test of fit than BFI or RNI because it does not exceed the 0 to 1 range. CFI values of greater than .9 are generally considered to indicate acceptable levels of model fit.

Because CFI has an upper bound of 1, CFI has been shown in many instances to be a slightly better (more efficient) index than RNI. As some authors have pointed out, however, RNI's ability to exceed 1 may provide useful information when the sample size is small and the researcher is examining nested models. Since CFI is constrained to the upper limit of 1, it can fall short in uncovering the magnitude of difference between models under the aforementioned conditions. It should be noted that many argue that almost all data is nested; not withstanding these conditions, CFI is still one of the best incremental fit indices available due to its efficiency (Goffin, 1993).

RMSEA

The root mean square error of approximation (RMSEA) is one of the most recently proposed tests of model fit (Steiger & Lind, 1980, as cited in Quintana & Maxwell, 1999). Because the value of RMSEA is less affected by sample size than is chi-squared, the RMSEA, like other alternative fit indices, has more descriptive value than chi-squared across various sample sizes. Interpretation of RMSEA values is often considered according to the following: 0 = perfect fit; <.05 = close fit; .05 to .08 = fair fit; .08 to .10 = mediocre fit; >.10 = poor fit (Byrne, 1998). RMSEA has been seen as a better indicator of fit than RMR (root mean square residual), an earlier model fit index upon which RMSEA is roughly drawn.

RMSEA's greatest strength is its ability to outline a confidence interval around its calculated value. Because the RMSEA's distribution values are known, a confidence interval around the point estimate of the RMSEA can be constructed to indicate the level of its precision. Using this confidence interval, evaluating the null hypothesis can be examined more precisely. In using these confidence intervals, a null hypothesis (H0: not a close model fit) could be rejected in favor of accepting the alternative (HA: close fit) if the entire range of the confidence interval is less than .05. So, like the chi-squared statistic, it is possible to use RMSEA to evaluate the null hypothesis that a model fits the data exactly. Unlike the chi-squared statistic, RMSEA is less affected by sample size problems.

On the negative side, relatively little information is currently available on the performance of RMSEA when data are nonnormal. What information that is available suggests that RMSEA may perform less optimally when there are large sample sizes and relatively small degrees of freedom (Quintana & Maxwell, 1999).

The development of procedures for calculating confidence intervals and for testing null hypotheses are important milestones in the development of SEM fit indices. While confidence intervals are likely to be produced for other fit indices (some of which may prove to be better able to test models with smaller degrees of freedom), RMSEA's value is in its originality.

Parsimony ratios

As mentioned above, the chi-square benchmark has been questioned by some researchers as it will almost always achieve statistical significance as sample size and degrees of freedom increase. Therefore, it has been suggested that the ratio of chi-square to degrees of freedom be examined as an alternative realistic indicator of fit when sample sizes and/or degrees of freedom are large (with a ratio of 5:1 to 2:1 considered to be within the acceptable range) (Tanaka, 1993). Various parsimony-weighted fit indices have been proposed (of which the former is one). These fit statistical weights, which range up to one and down to zero for just-identified models, are multiplied times indices such as NFI (or χ^2), to take model complexity into account and reward models that estimate fewer parameters, for the sake of parsimony.

Local Fit

Overall model fit may be poor, but there are other alternatives for determining the value of a model. Specifically, local model fit focuses on the value of individual parameters in the model. There are two ways of evaluating parameter estimates. The main test of parameter significance is a z-test for estimated individual paths (which should be > 1.96 at the .05 level of statistical significance). Additional support regarding local fit is indicated when significant paths are found to be in the hypothesized direction, and the magnitude of the item loadings is greater than .45 (parameter estimates or standardized regression weights) (Bentler & Wu, 1983; Jöreskog & Sörbom, 1989).

Modifications and Critical Ratios

There are several different ways to address model misspecification. Deleting statistically nonsignificant parameters increases degrees of freedom and can result in a more parsimonious model (simpler models produce better overall indices of fit). Modification indices (MI) are used to determine if additional "tweaking" of the specified model would result in a better overall model fit. MIs indicate if parameters that are "fixed" or constrained would be better off "freed" or estimated by indicating a possible decrease in χ^2. Researchers sometime consider double digit (10 or higher) MIs indicate parameters that would be best to free, however, this is not a stringent rule and should be handled on a case to case basis. The Wald test is an example of a critical ratio in use. The Wald Test (or z test as mentioned) is used to determine whether free parameters should be fixed. The Wald Test tells researchers possible parameters that should not be estimated (if absolute value of greater than 2 indicates a ripe parameter was correctly estimated). Importantly, all model specifications searches are tentative and should not be considered as true fits to the data. Original theory should drive decisions of model fit rather than post hoc model "tweaking." However, model specification searches may lead to better specified models in the future research

Summary

Researchers may be attracted to employ SEM because of its many benefits, yet they may lack the awareness regarding theoretical background of the multiple fit statistics thus limiting their decision about model fit. Researchers must determine whether they are interested in testing the null hypothesis, absolute fit, or incremental fit. In addition, researchers should be aware of the shortcomings of different fit statistics and how their model may lessen the applicability of specific fit statistics. While numerical cut-offs have been provided for evaluating fit statistics it

behooves researchers to investigate further and not blindly accept the criteria provided in textbooks as the popularity of SEM does not appear to be waning. In most cases, researchers should involve multiple criteria when evaluating their model fit.

References

Bentler, P. M., & Bonett, D. G. (1980). Significance test and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Bentler, P. M., & Wu, E. J. C. (1993). EQS/Windows user's guide. Los Angeles: BMDP Statistical Software.

Bollen, K. A., & Long, J. S. (Eds.). (1993). Introduction. Testing structural equation models (pp.1-9). Newbury Park, CA: Sage Publications, Inc.

Byrne, B. M. (1998). Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming. Mahwah, NJ: Erlbaum.

Fan, X. & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. Educational and Psychological Measurement, 58, 701-735.

Goffin, R. D. (1993). A comparison of two new indices for the assessment of fit of structural equation models. Multivariate Behavioral Research, 28, 205-214.

Henson, R. K. (1999). Multivariate normality: What is it and how is it assessed? In B. Thompson (Ed.) Advances in social science methodology. Standford, CT: JAI Press.

Hu, L. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), Structural equation modeling: Comcepts, issues, and applications (pp.76-99). Thousand Oaks, CA: Sage Publications, Inc.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-202.

Jöreskog, K. G. & Sörbom, D. (1989). LISREL 7: A guide to the program and applications (2nd ed.). Chicago: SPSS.

La Du, T. J., & Tanaka, J. K. (1995). Incremental fit index changes for nested structural equation models. Multivariate Behavioral Research, 30, 289-316.

MacCullum, R. C., & Hong, S. (1997). Power analysis in covariance structure modeling using GFI and AGFI. Multivariate Behavioral Research, 32, 193-210.

Quintana, S. M., & Maxwell, S. E. (1999). Implications of recent developments in structural equation modeling for counseling psychology. The Counseling Psychologist, 27, 485-527.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.). Testing structural equation models (pp. 10-39). Newbury Park, CA: Sage Publications, Inc.